Argonne NATIONAL LABORATORY | Feature

Cloud computing helps decode German *E. coli* strain

When a nasty strain of *E. coli* flooded hospitals in Germany this summer, it struck its victims with life-threatening complications far more often than most strains—and the search for explanation began.

Over a feverish weekend after the rogue bacterium's genome was sequenced, scientists from all over the world submitted the *E. coli* genome to rounds of rigorous study. Thanks to a unique Argonne-developed computer program and cloud computing testbed, researchers mapped the strain's genes—and came a little closer to understanding the bacterium's secrets.

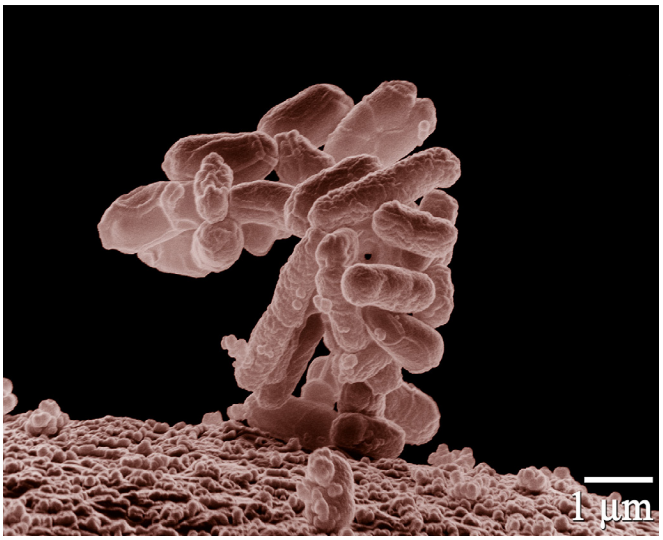
A team of Argonne scientists developed the Rapid Annotation using Subsystems Technology (RAST) program in 2007. The program, which is free and open to any scientist, is designed to make sense of the jumble of letters that makes up an organism's DNA.

A genome is a long, incomprehensible string of letters in a four-letter alphabet: G, A, T, C. Sections of the string are divided into genes. Each one describes how to build a protein, and proteins build all of the parts of the cell.

"If we can figure out what DNA codes for which protein, and what that protein does, then we can look at any bug and have an idea of what it can do," explained Ross Overbeek, an Argonne computer scientist who helped design RAST.

"For example, bugs with multi-drug resistance often turn out to have little pumps that drain the drug out of the cell as fast as it comes in," Overbeek said. "Once you know what those pumps look like, you can think about how to get around them."

RAST matches sections of the new string with its enormous catalogue of previously sequenced genes and proteins. At the end it spits out an annotated genome with a sort of "Cliffs Notes" to the organism's probable genes and proteins.



There are thousands of E. coli strains, but programs like the Argonne-developed RAST can help researchers make sense of a particular strain's genome. Photo credit Eric Erbe, digital colorization by Christopher Pooley, both of USDA, ARS, EMU.

When scientists announced they had sequenced the genome to the *E. coli* strain that plagued Europe on June 3, researchers from around the world began sending versions of the genome to RAST for annotation. They wanted to compare the new strain with past strains to tease out its origins and vulnerabilities.

"Genomes can vary even within a strain," Overbeek said. "You can get slightly different genomes in the same outbreak, even from the same patient. You compare genomes to see how the organism is mutating even as it's wreaking havoc."

RAST servers were already overwhelmed by a flush of

—*continued*

genomes and the new requests began to pile up—reaching more than 200 genomes an hour at one point. Its operators wanted to prioritize the *E. coli* work, so they turned to a resource designed for just such a possibility.

Magellan is a DOE test cloud computing project designed to boost research by making additional servers available on demand for scientific computing. The program, partially funded by the Recovery Act, has two sites— one at the Argonne Leadership Computing Facility and one at the National Energy Research Scientific Computing Center in California—but is designed to give researchers across the nation access to computing power in times of need.

The Argonne team duplicated the RAST server on Magellan, rapidly increasing the available computing power. "Our system is designed to use clusters, so we engineered it so that a piece of Magellan became part of the cluster that we use for RAST," explained Bob Olson, an Argonne computer scientist who maintains RAST.

It worked—so well that even more submissions poured in. Argonne and Virginia Tech teams worked around the clock that weekend to keep the servers running smoothly.

"They found exactly what they were looking for," Overbeek said. "The difference between this new strain and older ones came down to just a few genes. Apparently, the new strain included a combination of virulence factors present in other studied strains."

The operation was a perfect case for Magellan, Olson said, because each genome submission is an independent problem. Simply adding more processors to handle the extra jobs is easy—unlike many other computations, which often must solve successive problems; processors must wait to start their jobs until another processor finishes.

Overbeek remembered the early days of annotating genomes in the mid-1990s, when it took four or five scientists more than a year to analyze just one genome. "Now we can spit them out in a few hours," he said, and the team has already tested the next generation of RAST— a version so fast that it cuts the time for annotating a typical E. coli genome from eight hours to just 15 minutes.

"RAST is really revolutionary," Overbeek said. "It's turned a problem that used to be insurmountable into one that is trivial."

There is even an iPhone app to submit and receive genomes from RAST servers.

Developed at Argonne, RAST is funded through the National Institutes of Health and run by the Pathosystems Resource Integration Center (PATRIC) at the Virginia Bioinformatics Institute. PATRIC keeps a publicly available database of sequenced genome information.

The U.S. Department of Energy's Argonne National Laboratory seeks solutions to pressing national problems in science and technology. The nation's first national laboratory, Argonne conducts leading-edge basic and applied scientific research in virtually every scientific discipline. Argonne researchers work closely with researchers from hundreds of companies, universities, and federal, state and municipal agencies to help them solve their specific problems, advance America's scientific leadership and prepare the nation for a better future. With employees from more than 60 nations, Argonne is managed by UChicago Argonne, LLC for the U.S. Department of Energy's Office of Science.

By Louise Lerner.

NEWS CONTACT ▸ Louise Lerner | 630-252-5526 | media@anl.gov | Media Relations | www.anl.gov